Analysis of an Oral Skills Testing Instrument in a Teaching of English Undergraduate Program

John Edison Carvajal Arenas

Luisa Fernanda Castaño Ocampo

Universidad Tecnológica de Pereira

Facultad de Bellas Artes y Humanidades

Licenciatura en Bilingüismo con Énfasis en Inglés

Pereira

2017

Analysis of an Oral Skills Testing Instrument in a Teaching of English Undergraduate Program

John Edison Carvajal Arenas

Luisa Fernanda Castaño Ocampo

Trabajo de grado presentado como requisito parcial para obtener el título de Licenciado en Bilingüismo con Énfasis en Inglés

Asesor: Daniel Murcia Quintero

Universidad Tecnológica de Pereira

Facultad de Bellas Artes y Humanidades

Licenciatura en Bilingüismo con Énfasis en Inglés

Pereira

2017

TABLE OF CONTENTS

Abstract

The purpose of this research project was to delve into the reality of the language testing practices implemented in the Colombian EFL context. This was achieved through the analysis of a testing instrument used to measure learners' oral skills in one of the courses offered by a teaching of English undergraduate program. The language test usefulness model proposed by Bachman and Palmer (1996) was taken as main point of reference in the development of this study. Data was gathered through instruments such as interviews with the course's professor and students, observations, and a test analysis chart. This information was analyzed to determine how test usefulness qualities were evidenced in the test. As a result, construct validity and authenticity were more prominent than reliability, interactiveness, practicality, and impact. Additionally, elements related to test design and implementation, such as the importance of test objectives and scoring methods, were found.

Keywords: language testing, analysis of language tests, test usefulness, oral skills test

Resumen

El propósito de este estudio fue ahondar en la realidad de las prácticas evaluativas implementadas en el contexto de la enseñanza del inglés como lengua extranjera en Colombia. Ésto se logró a través del análisis de un examen que mide las habilidades orales de estudiantes pertenecientes a un programa de pregrado en lengua inglesa. El principal punto de referencia para el desarrollo de este estudio fue el modelo propuesto por Bachman y Palmer (1996) para la evaluación de exámenes de lengua. Se realizaron entrevistas con el profesor y estudiantes del curso en el que se llevó a cabo el estudio, así como observaciones durante la administración del examen, y la adaptación de un instrumento para el análisis del mismo. La información obtenida fue analizada para determinar cómo se evidenciaban ciertas cualidades que determinan la utilidad de un examen de lengua. Como resultado, las cualidades de evaluación más visibles fueron la viabilidad y autenticidad del examen, mientras que su fiabilidad, interactividad, practicidad e impacto fueron percibidas en menor medida. Además, se encontraron otros elementos relacionados con el diseño e implementación de exámenes, como la importancia de los objetivos de evaluación y métodos de calificación.

ACKNOWLEDGEMENTS

To our parents, friends, mentors, advisor, and anyone else to whom this may concern:

Our deepest gratitude.

Sincerely,

Luisa & John

1. STATEMENT OF THE PROBLEM

Language testing has played a significant role and continues to gain prominence in the EFL field. Throughout the years, the way learners' abilities are measured has been subject of research and continuous improvement by language testing experts all around the globe, and Colombia is not oblivious to this matter. In fact, the 2017 Language Testing and Research Colloquium (LTRC) held by the International Language Testing Association (ILTA) took place in Bogotá. Both the association and the colloquium are international stances that promote the inquiry and enhancement of language testing practices around the world. They provide a space for researchers, scholars, and practitioners on the field to meet, share, and discuss new ideas and research findings related to language testing. By taking part of this event, Colombia assures its current interest in the EFL field and commitment regarding the research and improvement of language testing practices, which, in turn, supports the endeavor towards its national bilingual policy.

The country's efforts to develop learners' communicative skills in the target language involve language assessment and, therefore, language testing as fundamental elements in the consolidation of such policy. This is clearly stated in the set of curriculum guidelines proposed by the National Ministry of Education, "*Lineamientos Curriculares para el Área de Idiomas Extranjeros en la Educación Básica y Media*", along with a set of standards known as *"Estándares Básicos de Competencias en Lenguas Extranjeras: Inglés"*. The former being a model for foreign language teaching in the country's educational system, which presents a section completely devoted to the evaluation and assessment process, including its purpose, development, and impact, as well as the importance of its appropriate implementation. And the latter, being more specific parameters for the design and implementation of English lessons in terms of what learners should

be able to know and do according to each grade, which consequently affects the language testing process.

Altogether, these undertakings ought to be evaluated to have a notion of their effectiveness and make informed decisions based on the results of their implementation. The Colombian educational system evaluates such outcomes through a test known as *"Examen Saber 11",* which is implemented at the end of the high school stage. This test measures students' abilities in fundamental areas of knowledge, including English. It is worth mentioning that all three curriculum guidelines, basic standards, and standardized test are aligned with the proficiency levels proposed by the Common European Framework of Reference for Languages (CEFR), an internationally recognised framework used by organisations all over the world as a reliable benchmark of language ability. By taking these international parameters as reference for the development of the various initiatives aiming at the national bilingual policy, Colombia nears its main objective of instructing citizens able to communicate in English, so that the country can be involved in processes of universal communication, global economy, and cultural overture.

At the core of this national endeavor, English language teachers play a significant role, and therefore, their instruction in the teaching and assessment of the target language is a matter of concern within the EFL Colombian context. This issue has been addressed by researchers such as Cohen and Fass (2001), Frodden, Restrepo, and Maturana (2004), and Pineda (2014). They conducted studies in English programs from different Colombian universities, in which they identified, described, and analyzed teachers' practices in terms of language assessment and oral instruction. In the end, they all arrived at the same conclusion: the imperative need of teacher training in these areas. This is in itself a decisive aspect that requires careful attention in the process to consolidate the national bilingual policy due to the fact that its successfulness depends to a

significant extent on the quality of the professionals in charge of instructing learners. However, this only sets the path for further investigation and reflection about the role of language assessment in the country's EFL field.

With this in mind, the reason to conduct this study stems from the fact that no research was found in regards to the assessment practices implemented in the Teaching of English undergraduate program where it was developed. Besides, all of the studies previously mentioned were exclusively carried out in English institutes belonging to Colombian universities. This leads to question the way in which future English teachers are currently assessed, even more since they will face at a certain point in time the demanding task of not only teaching the target language but of eventually assessing students' performance. Such concern fits right into the national endeavor to enhance English teaching and learning in the country and the EFL field's interest to research on language testing. This research project, then, aims at a better understanding of the way future English teachers themselves are assessed.

This will be done through the analysis of one of the instruments used to test learners' skills in a Teaching of English undergraduate program. The testing instrument analyzed is part of a subject called Academic Discourse, which is completely endeavored towards the enhancement of future teachers' oral skills. Focus on oral ability stands from the fact that the practice of testing second language speaking is the latest subfield of language testing (Fulcher, 2014), and that speaking skills are usually left behind in terms of efficient and reliable assessment procedures in spite of holding such importance in the development of the communicative competence (Cohen and Fass, 2001; Pineda, 2014). Furthermore, it is worth mentioning that emphasis on this specific skill strives from the particular interest of the project's researchers and their experiences as English teachers to be.

This analysis will be based on Bachman and Palmer's model for language test usefulness proposed back in 1996, which presents a set of qualities that determine to what extent a test accomplishes its intended purpose. It will be carried out in order to raise awareness of what is currently done, what works and what does not, and what can be done in order to enhance language assessment practices in the already described scenario. Therefore, the main research question that frames this study is:

- What can be evidenced, in terms of test usefulness, through the analysis of an instrument used to measure learners' oral skills in the Academic Discourse subject of a Teaching of English undergraduate program?

2. CONCEPTUAL FRAMEWORK

The main theoretical concepts that support this research study will be defined in this section. The first construct that will be addressed is language testing and its role in the English language teaching field, along with a classification of the different types of tests found in such context. The second concept centers on testing speaking since this research aims at the analysis of a test that measures said skill, which happens to be one of the less explored in language testing. The last construct covers the model of test usefulness proposed by Bachman and Palmer (1996); being its qualities the reference frame for the analysis conducted in this research project. In order to have a clear understanding of these concepts, they are defined as follows:

2.1. Language Testing

Language testing is regarded as a crucial practice in the English teaching and learning field. Anyone who has ever been involved in managing another language knows that tests are a recurrent theme, and that they come of all types and forms, with different purposes depending on each context and testing situation. Language tests are the instruments used to measure learners' abilities in the foreign language and play a significant role when making decisions in regards to the teaching and learning practices of individuals, programs, and even institutions. Bachman and Palmer (1996) highlight the importance of language tests by referring to them as "a valuable tool for providing information that is relevant to several concerns in language testing" (p. 8). Such information includes evidence of the teaching and learning practice that contributes to the enhancement of teaching programs themselves; relevant information for making decisions about individuals, such as determining the best ways to present the target language to students based on their specific needs, strengths , weaknesses, and interests; or assigning grades to represent learners' linguistic achievements. This means that, in effect, all language teaching and learning related programs must

deal with testing in any of its forms, which implies that language teachers need to develop skills that will allow them to make the right decisions when it comes to select, plan, construct, or implement their own language tests.

Carr (2011) concurs with Bachman and Palmer's position towards language testing. He reassures the importance of this practice for language teachers, who often face the demanding task of developing or choosing tests or other forms of assessment, either for a whole institution or a class, with little or no training on how to do it. This author also makes the assumption that language tests are tools used to do a determined job, for a particular reason, with a specific purpose, which most of the times involves making decisions about learners or other people related to the language testing process. In order to have a clearer view of these particular reasons for developing language tests, Carr (2011) presents the following classification in terms of test purposes and types:

Admission test. These types of tests are used to decide if a learner should be admitted to a certain program or not.

Placement test. These are intended to classify learners according to their language proficiency level, placing them in the appropriate stage to continue developing their abilities.

Diagnostic tests. These are carried out as a way to identify possible gaps in language users' command of the target language, and function as a starting point to make decisions about their learning process. Sometimes this kind of information is gathered from both admission of placement tests.

Progress tests. These are intended to assess how well students are mastering the content or objectives of a course as a means of monitoring or following their performance or progress. An example of these types of tests would be quizzes.

Achievement tests. These instead, are used to identify how well students have mastered the content or met the objectives of a particular course. Both progress and achievement tests can be used to make decisions about individuals or the teaching program itself. To what extent a progress test can become and achievement test or the other way around will depend upon the use that such test is given.

Proficiency tests. These tests certify learners' overall language ability and are commonly used as requirements to access education or job opportunities that demand certain command of a target language. International examinations such as the IELTS and TOEFL tests are examples of these.

Screening tests. These would be proficiency tests used to make selection decisions, such as qualifying for a particular job.

Apart from the purposes and types of tests mentioned above, Carr (2011) states that tests can be seen from other perspectives, such as the interpretation of results, the different things test takers have to do during the test, and the scoring process. These are described as follows:

Norm-referenced tests (NRTs). When test takers' results are compared to how well others did on the same test.

Criterion-referenced tests (CRTs): In this case language users' ability is assessed in terms of standards, objectives, or other criteria, and not compared to others' results.

Summative assessment. This type of assessment is usually given at the end of a unit, course, program, etc., and provides information about how much students learnt. It is closely related to achievement tests, and in fact, most achievement testing is largely summative, and summative testing usually aims to assess learner achievement.

Formative assessment. This type of assessment takes place during the process of learning something and is used as a means to monitor learners' progress. It is closely related to progress assessment. In order to make a distinction between summative and formative assessment, it is useful to remember that the former is used to sum up how well learners did, and the latter is used to shape or form what is being taught

Objective test. Tests that can be scored objectively through the use of selected-response questions, such as multiple-choice, true-false, or matching questions.

Subjective test. This one involves human judgement in the scoring process, just as in most writing and speaking tests.

Direct test. These tests require test takers to use a specific ability; e.g., speaking in a speaking test; writing in a writing test.

Indirect test. These tests aim at assessing a productive skill through related tasks that do not require any speaking or writing.

Semi-direct tests. These are usually implemented for speaking tests that demand test takers recording their speech instead of talking directly to a human interlocutor.

Discrete-point test. These type of tests use separate unrelated tasks or questions to assess one "bit" or language ability at a time; e.g., multiple choice. It is very useful for testing very specific areas of language, such as grammar.

Integrated test. These type of tests require learners to utilize multiple elements of language ability in more life-like tasks; e.g., taking notes of a lecture and then writing to write a summary. Since these tests resemble real life, they more communicative language use.

Independent speaking and writing tasks. These would be the most common approach in speaking and writing assessment, and consists of test takers reacting to their interlocutors' prompts.

Performance assessment. It involves the actual performance of relevant tasks. The strong sense of language performance assessment is concerned with the completion of a task similar to those found in a real life scenario, and for which linguistic accuracy only matters if it interferes with the learner's performance. And the weak sense of language performance assessment deals with the level of the language used to perform the task. Its main purpose is to elicit a sample of language to be evaluated, and the learner's completion of the task comes in second place.

Being able to identify what type of test is implemented in a particular testing situation proves to be of great value when attempting to analyze a test itself, as it is intended to happen in this research project. As Bachman and Palmer (1996) and Carr (2011) stated, the most important quality of a given test is its purpose, which is closely related to the type of decisions it is used to make. However, there is more to language testing than having a clear purpose, and it is precisely the part that requires more attention from language teachers. This following stage in the testing process involves deciding on the types of tasks learners must perform and the way those tasks are to be scored. Such aspects change depending on which specific language skill one is attempting to test. In this case, the main focus of this research study is on a subfield of language testing; this is, testing speaking. Therefore, it is precisely this concept the one to be defined next.

2.2. Testing Speaking

As stated in the Common European Framework of Reference for Languages (2001), all language tests are considered instruments to assess learners' proficiency. Speaking, however, proves to be a difficult skill to measure since it deals with more than the mere linguistic command of the language. In fact, in speaking tests, learners' individual characteristics' influence is more evident due to their immediate nature. Factors like students' personality, background, interests, proficiency level, and learning styles emerge and shape their potential outcomes. Aside from these,

there are also other elements such as the learning environment, testing criteria, syllabus specifications, and rater's role, to name a few, that come into place during the testing process. Luoma (2004) concurs with the importance and impact of these aspects in the assessment of learners' spoken performance. She asserts that, through speaking, individuals reflect signature features of their own, and that testing outcomes depend to a great extent on the context, tasks parameters, implemented criteria, and the relationship between the interlocutor and the examinee.

Another key element in the speaking testing process is the type of tasks learners are exposed to. Luoma (2004) addresses this issue as she describes and categorizes speaking tasks in terms of communicative functions and types of talk. Both involve actions such as describing, narrating, instructing, comparing, explaining, justifying, predicting, deciding, reporting, asking, agreeing, disagreeing, suggesting, requesting, warning, socializing, summarizing, structuring discourse, and repairing communication. Fulcher (2014), on the other hand, highlights the importance of developing speaking tasks that replicate a real life usage of the target language. He also provides a framework for describing speaking tasks and categorizes them in terms of task orientation, interactional relationship, goal orientation, interlocutor status, topics, and situations. These categories, then, allow test designers to develop appropriate tasks for different specific purposes. Finally, both authors concur with the notion of implementing rating scales as measuring instruments in order to avoid subjectivity within the testing process. These instruments are also known as "scoring rubrics" or "proficiency scales" and consist of a series of grading criteria that includes quantitative and qualitative descriptors of what test takers are expected and able to do with the language. They argue that these rubrics can be adjusted in accordance with aspects such as test purpose and target audience, that they provide a clear set of criteria for the development of tasks and allow testers to have expected outcomes for their scoring procedures.

The practice of testing speaking poses a variety of challenges as there are many aspects to take into consideration in both the design and implementation processes. These aspects have a significant impact on learners' performance and, consequently, on test scores and the information drawn from these. In that sense, task design and scoring procedures play key roles in the construct of testing speaking. However, all of this is complementary to the particular testing situation. In fact, what determines the way a test is designed and implemented is its specific purpose. For that reason, the extent to which a test accomplishes its intended purpose becomes a matter of concern. This issue, then, is portrayed in the next concept.

2.3. Test Usefulness

Language tests are intended to measure learners' abilities to use the target language, so their ultimate purpose is to define what language users can do and how well they can do it. Bachman and Palmer (1996) state that the most important aspect in the design and development of a test is that it fulfils its designated purpose. They describe this as test usefulness and present a model with a range of qualities that determine to what extent a test accomplishes such characteristic. These qualities are defined as follows:

Reliability. This is defined as consistency of measurement. In other words, a test is considered reliable when its scores show consistency within and among groups of examinees and raters.

Construct validity. This refers to the justification of the interpretation given to tests scores. Therefore, construct validity is evidenced when there is a coherent relation between the definitions of what is measured for example, the course syllabus, the testing tasks, and the tasks conducted in the target language use context.

Authenticity. This is the equivalence between language tasks that learners might encounter in tests, and the ones they might face in a real-life scenario.

Interactiveness. This relates to the way individuals' characteristics are engaged in a test task. In other words, the interaction test takers have with a given test. Clear layout and instructions, appealing content and tasks, and how learner friendly the test is, are all aspects to take into consideration.

Practicality. This quality involves the resources utilized to design and develop a test and its success relies on the availability of said resources, which not only refers to materials, but also to time constraints and other circumstances connected to the individuals taking part.

Impact. With the implementation of tests come a series of implications that resonate within various groups of individuals, operating at two particular levels: At a micro level, concerning the individuals that are directly affected by the test implementation, such as the examinees, and at a macro level, concerning administrators and the specific educational system.

Although Bachman and Palmer proposed these concepts back in 1996, they are still taken as a point of reference in the English language testing field. Carr (2011), for example, restates these qualities as fundamental aspects in the design and analysis of language tests, making especial emphasis on the fact that each quality is given a different degree of importance depending on the testing situation without disregarding the rest.

These qualities have also been explored in the EFL Colombian context by authors like Frodden et al. (2004) and Pineda (2014) in their studies on assessment instruments and oral English language performance respectively. Frodden et al. (2004) found that practicality and reliability are the core qualities in the assessment procedures of two national universities as there is always a need to save time, materials, and efforts while getting consistent results. Similar to this research, Pineda (2014) emphasizes the importance of practicality. She describes the benefits of utilizing rubrics in terms of saving time for teachers who usually face the overwhelming task of grading

loads of papers. These studies set the path for delving into the reality of language testing practices in our own context, specifically the ones involving oral skills assessment.

Oral skills testing instruments are shaped by several different factors including their purpose, task design, audience, context, grading procedures, to name a few. The concept of testing qualities, on the other hand, aids their analysis in terms of design and implementation. The fact that oral performance is a highly subjective skill to assess, and that there is a need to deepen in the current practices implemented at the UTP's Teaching of English program, offers a golden opportunity to analyze the instruments used to test oral skills. This could provide a better understanding of what is done, what works and what does not, and what could be done in order to raise awareness on how to enhance such practices in the established scenario.

3. LITERATURE REVIEW

Assessment in a foreign language classroom has become a matter of concern within the English teaching field; ranging from traditional to alternative procedures, and comprising the well-known communicative skills of listening, reading, speaking and writing as core focus. Bearing this in mind, speaking as a productive skill is fundamental in foreign language teaching and learning, and also one of the most subjective competences to measure. To address this issue, various researchers in Colombia, including Frodden, Restrepo and Maturana (2004), Cohen and Fass (2001), and Pineda (2014), have carried out studies where they identified teachers' and learners' beliefs about assessment, in addition to criteria and instruments commonly used for this purpose in the foreign language classroom. These investigations have drawn conclusions such as the reality of assessment in our context and the imperative need of teachers' training in oral instruction and assessment, which are related, valuable, and enlightening to this research that arises from the need to have a better understanding of the qualities that determine the usefulness of the instruments implemented to test learners' oral proficiency in a Teaching of English undergraduate program.

Shedding some light on the broader topic of foreign language assessment, Frodden, et al. (2004) portray some findings on an investigation carried out to analyze and rectify teachers' assessment processes in the foreign language classrooms of two Colombian universities. Following a collaborative action research approach, twelve English teachers and five French teachers voluntarily participated in this study by providing the instruments used to assess their learners' oral production, responding to interviews about their assessment practices in the foreign classroom, and being part of awareness workshops conducted to delve into this matter for the sole purpose of agreeing on a refined assessment methodology. This way, researchers categorized all

the information collected in order to design a chart that facilitated the process of analyzing assessment instruments and tasks.

As a result, hard types of assessment were acknowledged to be the favored ones by the participating teachers in their pedagogical practices, as well as practicality and reliability as core qualities in their assessment procedures. Frodden, et al. (2004) describe hard types of assessment as the conventional assessing techniques that focalize on results rather than the whole process, such as quizzes and exams. These instruments are thought to be objective, precise and reliable since they include items with systematic grading scores, like true or false, multiple choice, and matching exercises that make them easier to design and grade. Therefore, it has been highlighted that practicality and reliability are the guiding principles for teachers in their assessment practices as there is always a need to save time, materials, and efforts while getting consistent results. Bearing this in mind, skills such as listening, reading and writing were determined to be the preferred ones to be developed in a language class, leaving behind oral instruction and therefore its assessment practices.

Deepening precisely into the implications of speaking skills instructions and their assessment, Cohen and Fass (2001) compared facilitators and students' beliefs about the development of speaking skills in an EFL scenario with their current practices. This study was conducted in a Colombian university English program, with the participation of 40 teachers and 63 random students from various proficiency levels, from beginners, intermediate, and advanced courses. It involved teachers and learners' questionnaires and classroom observations in conjunction with the identification of three main aspects: participants believes, materials used and assessment procedures. Each aspects was addressed by a different group, being "the Beliefs

Group", "the Materials Group", and "the Assessment Group"; this allowed to target and deepen in each of the aspects mentioned in an independent and clear way.

Researchers concluded that beliefs and reality differed, that textbooks were usually adapted to meet learners' needs, and that teachers needed to be trained in oral assessment practices. Regarding the differences between beliefs and reality, it was determined that teachers and learners had opposite conceptions about talking time in the classroom, with the former in favor of student active participation, and the latter quite reticent towards it. Then, believes were confronted with reality by recorded classroom observations, leading to identify that classes were indeed teacher centered; hence, contradicting teachers' perceptions about communicative classes. Moreover, teachers usually adapted the institution's textbook so as to make it applicable to their classes but without modifying its non-communicative nature. Assessment also proved to concentrate on pronunciation and grammatical accuracy, leaving aside aspects of spoken language, such as fluency and the ability to make oneself understood. On this basis, it was finally concluded that teachers needed appropriate training to achieve real communicative instruction and valid assessment of oral skills.

Addressing this specific issue of ensuring effective assessment practices, Pineda (2014) draws attention to the implications of designing and implementing a rubric to enhance oral production assessment in a language institute of a Colombian university. The research was carried out with the participation of 39 teenage students from four level-one English courses and their respective teachers. The investigator, as coordinator of the language institute where the investigation took place, realized her teachers' ongoing concern of improving speaking testing conditions and decided to research on the matter. Initial interviews and surveys were administered to identify the aspects they most needed to reflect and train on regarding the assessment of

students' oral performance. After this, Pineda (2014) brought up a proposal for testing oral skills by the means of rubrics, taking into account all the information collected, and later created an oral task with its correspondent rubric for teachers to implement, and measure its usefulness in the assessment process.

Similar to Frodden et al. (2004), Pineda (2014) emphasizes on the importance of practicality, but in this case, describing the benefits of utilizing rubrics in terms of saving time for teachers, who usually face the overwhelming and never ending task of grading loads of papers. As main conclusions, the author highlights three aspects: First, the making of a rubric for oral assessment might be time consuming, but becomes a time saver when grading; second, rubrics grant objectivity to the assessment procedure since it establishes clear criteria in the levels of performance; and lastly, teachers need to have more training in assessment before measuring learners' speaking skills. The author stresses on the difficulty of designing these kinds of instruments for the first time, yet highlights their potential long term benefits, not only as time savers in evaluation, but also as opportunities to provide meaningful feedback to students' performance. Besides, it enables teachers to have an expected criteria for students beforehand while granting the possibility of providing fast and thorough feedback considering that all grading aspects are segregated into well-defined dimensions, making it fair and transparent. All this, reduces the lack of consistency commonly found when testing oral skills, especially in communicative contexts.

All of the authors mentioned, address the importance of assessment within the foreign language field, taking into account background information about the development of oral instruction and assessment in the Colombian context, insights on its reality supported by teachers' current practices and perceptions, and ways of enhancing its process. They also came to similar

findings that revolve around four main assessment aspects that should be refined: construct validity, reliability, practicality and assessment training. The first refers to the interpretation of test scores, which is not possible without a coherent connection between instruments' tasks and learners' expected performance in real context. The second relates to consistency in measurement through the design and implementation of assessment instruments, minimizing variations of task characteristics that affect test performance. The third deals with the need of reducing time on the design, administration, scoring and analysis of tests' results to make teachers' duties more manageable. And the last one brings up the need of adequate assessment training for teachers; if they are clear about how they should test students' oral skills and thoroughly explain them what is required, it will strengthen both teaching practices and students' learning in such a way that testing will become fairer and more transparent.

All of these contributions aim at improving language teaching practices along with learners' satisfaction and success; they provide different perspectives, confirming their impact on the language teaching and learning context. Furthermore, comparing and contrasting the different factors that take part in the assessment process, including learners' , teachers' and even institutions' points of view grants a high level of reliability to the conclusions drawn from these studies. The various methodologies, instruments and theoretical support utilized throughout the investigations prove to be similar to one another, including previous diagnosis, collection and analysis of testing instruments, recorded classroom observations, interviews and surveys for both teachers and learners, among others. These similarities confirm their effectiveness and reliability for such kinds of investigations. Nevertheless, some inconsistencies were noticed in terms of participants and data collection since findings were generalized; the studies were only focused on

specific target populations, not only setting apart other potential and meaningful groups of interests, but also failing to make it explicit and clear on their papers.

Looking forward, assessment procedures in the foreign language classroom, especially in regards to oral skills, should be addressed as an overall, under the scope of various useful testing qualities, apart from practicality and reliability which are normally the privileged ones. By attempting to reinforce only a particular quality, teachers might neglect the others affecting the ideal balance that ensures testing successfulness. Nonetheless, to achieve such endeavor; appropriate oral instruction and assessment training is vital for all parts involved. Likewise, teachers, learners, and language institutions play an important role in the assessment scenario, thus, should work as a whole, connecting their strengths, weaknesses, needs and expectations, according to the different circumstances and characteristics of a domain that demands constant improvement and renewal.

This research project has been shaped by elements from the studies previously addressed due to the relevance of their contributions to the assessment practices in the ELT field. Frodden et al. (2004) supported the adoption of the test usefulness model proposed by Bachman Palmer (1996) for the development of the analysis conducted in this project. On the other hand, Cohen and Fass (2001) and Pineda (2014) contributed to the notion of focusing on oral skills since it is, as they assert, one of the less explored abilities in the EFL context. The investigations presented in this section were taken as points of reference not only from both topical but from a methodological stance; the inclusion of data collection instruments such as interviews, observations, and an analysis chart are an example of this. The development of this study is then another step towards the understanding of the language teaching practices implemented in the Colombian scenario, with the difference of being developed in a program that instructs future language teachers.

4. METHODOLOGY

4.1. Type of Research

This research study aims at analyzing one of the instruments used to test oral skills in a Teaching of English undergraduate program. As Merriam (2009) states, the best way to approach research in the education field is through a qualitative design that allows to have a broader insight about one's practice, which, in fact, can lead to its improvement. In this sense, the present study follows a qualitative research approach.

4.2. Type of Study

This study will be conducted as a descriptive case study since it can be considered an in-depth analysis of a bounded system characterized for being particularistic, descriptive, and heuristic (Merriam, 2009). This research project is narrowly focused within a specific context and setting; it is mainly descriptive as it involves the analysis of both objective and subjective data; and it is developed in an attempt to shed light upon the issue of language testing in a concrete EFL scenario.

4.3. Context

The current study will be carried out at a Colombian state university located in the municipality of Pereira, the capital city of the Department of Risaralda. The university where this research will be developed was given an eight-year high quality re-accreditation by the MEN in 2013. Based on the institution's 2016 statistics, there are 87 academic programs currently offered by the university, being 32 undergraduate programs and 55 graduate programs. Regarding student population and teaching staff, there is an overall of 16.816 undergraduate students and 1.895 graduate students, as well as 1.292 professors hired under various modalities. Framed within a

particular EFL context, this university is the only higher education institution in the region to offer a teaching of English undergraduate degree through in-person instruction.

This Teaching of English undergraduate program is the one in which this research project will be taking place. It was created in 2004 as an attempt to fulfil the institution's needs regarding the teaching and learning of a foreign language, in this case English, and address the national endeavor towards bilingualism. It is a five-year undergraduate program that currently comprises a community of 697 students and 35 professors. As stated in the Program's Educational Project (PEP), it is driven by a main pedagogical model, constructivism, which is enriched by different approaches regarding language teaching and learning, being: humanism, critical reflection, and content based instruction. It is also divided into the English -Spanish, Pedagogical-Linguistic, Investigative, Technological, and Intercultural areas. The English-Spanish area aims at fostering the development and improvement of the future English teachers' communicative competence in both target language and mother tongue. The course considered for the implementation of this research study, the Academic Discourse Course I, belongs to this very area.

4.4. Setting

The Academic Discourse Course I offered by the Teaching of English undergraduate program seeks to develop and improve learners' proficiency in oral production regarding a B2 English level; this, according to the qualitative aspects of spoken language such as range, accuracy, fluency, interaction, and coherence, portrayed in the Common European Framework of Reference for Languages (2001). This course also aims at developing learners' teaching competences concerning the target language oral instruction, as well as the competences regarding the future English teachers' academic discourse particular to their EFL context. The Academic Discourse I

course is assigned with three in-class hours and six hours of autonomous work and belongs to the 4th semester of the undergraduate program.

4.5. Participants

After having identified the context and setting in which this research project will be taking place, it becomes necessary to select a sample from which data will be gathered. Taking this into account, the sample selection of this study is based on the method of purposive sampling. Fraenkel and Wallen (2009) define this type of sampling as the one in which "researchers use their judgment to select a sample that they believe, based on prior information, will provide the data they need" (p. 99). Bearing this in mind, this research will be conducted in the Academic Discourse I course of the Teaching of English undergraduate program. This course is completely devoted to the development of learners' oral skills, which makes it the most suitable option for this study, whose main purpose is to analyze one of the instruments used to test learners' oral ability. It will also count on the participation of the course's professor and students, who are described as follows:

4.5.1 Students

This study will be carried out with the participation of students of the already mentioned Academic Discourse Course I. Based on the Teaching of English undergraduate program's curriculum, students enrolled in this course must have met the prerequisite of Intermediate English. This course is mixed-gender, and learners' ages go from 18 to 24 years old, with the exception of a particular student who is 53. Recalling Fraenkel and Wallen (2009), purposive sampling is "based on previous knowledge of a population and the specific purpose of the research" (p. 98); in this sense, students' sample was selected taking into consideration the professor's impressions of students' awareness and critical thinking about their learning process. This information was also confirmed by the professor who guided the previous Intermediate English course, which was the

prerequisite of Academic Discourse I. The participation of critical and willing students enriches the quality of the data collected and is aligned with the research study's main purpose, which requires to take an insightful stance towards language testing. As for sample size in descriptive studies, a minimum of 10% of the population is suggested, and it should be as large as researchers consider reasonable (Fraenkel and Wallen, 2009). In this case, the total population corresponds to 32 students, out of which a sample of 6 was purposively selected; approximately the double of the minimum suggested (18.75 %). This does not only enrich the quality but the quantity of the data collected since the larger the sample is the more reliable the findings can be.

4.5.2 Professor

The professor guiding this course is the one in charge of designing and implementing the instruments used to measure learners' oral skills, and is therefore regarded as one of the participants in this study. The Academic Discourse I course's professor has taught this course in previous occasions. Regarding his experience in the language testing field, he guides the Curriculum Design course, which has a language testing component, and he has participated in conferences, seminars, and workshops as speaker. The selection of the course's professor is also based on purposive sampling (Fraenkel and Wallen, 2009).

4.6. Researchers' Roles

During the development of this study, researchers will play different roles, among which are the design, adaptation, and implementation of instruments for further data collection and analysis. Researchers will conduct interviews for which they will design a form containing the main aspects to be addressed. On the other hand, during the implementation of the testing instrument, researchers will play the role of complete observers; this is, they will look for behavioral aspects that might go unnoticed through other data collection methods. As Merriam

(2009) highlights, interviews and observations are considered major means of collecting data in qualitative research studies. Additionally, researchers will adapt a checklist to be implemented for the analysis of the testing instrument.

4.7. Data Collection Methods

As a means to give response to the research question that guides this project, data were collected through three different instruments, being interviews, observations, and an analysis chart. These instruments were implemented at three different stages, before, during, and after the test took place, since the qualities to be analyzed could be evidenced throughout the whole testing process; this is, its design, development, implementation, and results. These data collection tools were applied according to the course's professor's and students' availabilities and test's timetable. Interviews and observations were conducted over a period of three weeks, and the analysis chart was filled from the beginning to the end of the testing process. Both researchers were present in every data collection stage so as to foster the reliability and validity of the data obtained. All of this was done in order to have a broader scope in regards to the information gathered and enrich the quality and quantity of potential findings. These data collection instruments are described as follows:

4.7.1. Interviews

Interviews are among the most common data collection instruments in qualitative research. In fact, Merriam (2009) argues that they are often the major sources of qualitative data. She describes them as conversations with a specific purpose, which is to gain access to information that cannot be acquired otherwise. These tools become fundamental when there is a need to observe behaviors, to interpret feelings and conceptions, or to obtain information regarding past events that cannot be replicated. Furthermore, Merriam mentions three types of interviews in terms of

structure, ranging from highly structured or standardized, semistructured, and unstructured or informal. She states that, in qualitative research, all three types can be interwoven in order to develop a more thorough data gathering process.

Taking this into account, and for the purpose of this study, instruments with these characteristics were implemented as a means to enrich the information collected in terms of quality and quantity. Two different interviews were utilized at two different stages for both the course's professor and students (See Appendices E1, E2, G1, and G2). Both researchers took turns as interviewers and observers in this process in order to gather as much information as possible, not only from participants' answers but from their behavior. Participants' availability and test's schedule were taken into consideration when planning the interviews agenda. This was done before and after the test took place, so as to have a notion of participants' beliefs and impressions in both phases. Participants were given the option to answer in either English or their mother tongue, Spanish, to avoid any misunderstanding and facilitate their interaction with the interviewer and the questions made. It is worth mentioning that these questions were formulated bearing in mind the qualities that determine the usefulness of a test (Bachman and Palmer, 1996).

4.7.2. Observations

As well as interviews, observations are considered a major source of data when developing qualitative studies. They are thought to be complementary to other data collection methods, and their main purpose is to obtain firsthand information drawn from the research context itself. This method is particularly useful for accessing information that underlies written and oral data forms since it allows to perceive behavioral aspects that contribute to the holistic analysis of the issue under study (Merriam, 2009). In order to carry out observations in this research project, researchers played the role of complete observers during the implementation of the testing instrument. An

observation chart was designed and utilized to facilitate field notes taking and their subsequent systematization and analysis (See Appendix F). This particular type of data collection was developed not only in the implementation of the test, but also at different periods in time, such as teaching, advisory, and feedback sessions, both before and after the test. The main objective of this observation process was to capture aspects that gave account of the test usefulness qualities proposed for the analysis of the testing instrument.

4.7.3. Analysis Chart

This data collection instrument is an adaptation of a checklist proposed by Bachman and Palmer (1996) for the evaluation of test usefulness (See Appendix D). Checklists are defined as formats that set specific parameters to be met as part of a particular task (Fraenkel and Wallen, 2009). However, using such tool to collect data in this research project was not feasible since the main objective is to analyze a test instrument based on certain criteria, not to evaluate if these criteria are met or not. In this sense, the original checklist comprises a set of parameters based on the qualities proposed for test usefulness, being reliability, validity, authenticity, interactiveness, practicality, and impact (Bachman and Palmer, 1996). These elements served as a point of reference regarding the aspects that this research project seeks to unfold. This analysis chart was filled throughout the whole testing process, adding data as it was evidenced in different stages from beginning to end.

4.8. Data Analysis

In this research project, three data collection instruments were implemented: Interviews, observations, and an analysis chart. The method selected to analyze the information gathered is known as Content Analysis. This is a method that consists mainly in the analysis of documents and human communications and has wide applicability in educational research. Fraenkel and

Wallen (2009) suggest some steps when using this type of analysis: Determine objectives, define terms, specify the unit of analysis, locate relevant data, develop a rationale, develop a sampling plan, formulate coding categories, check reliability and validity, and analyze data. The first six steps mentioned have been previously addressed in other sections of this paper; therefore, this segment focuses on the last three.

After implementing the instruments to collect data, the process of coding took place. To do this, it was necessary to go over the information gathered and find aspects that drew researchers' attention. These aspects were then systematized by using codes to identify patterns which were then classified into main categories. All the information collected was transcribed in order to facilitate the exercise of coding, finding patterns, and categorizing. Throughout the development of this research project, emphasis has been made on six main qualities that determine the usefulness of a test (Bachman and Palmer, 1996); therefore, all data was first analyzed in order to identify elements characteristic of each quality. All of this information was indexed in a data matrix with the test usefulness qualities as pre-established themes and a coding system to classify each piece of information based on its source. For instance, the code used for the interviews conducted before the administration of the test followed the format "*L55IBS5*"; "L55" stands for the transcript's line where information was found; "I" stands for "Interview", and "B" stands for the moment in which the interview took place in relation to the test, in this case "Before administration"; and "S5" stands for the type of participant, in this case "Student 5".

The following table shows a more detailed description of the coding system implemented in the data analysis process:

Table 1

Data Analysis Coding System

Code	Description	Example
L#	Transcript's Line	***L55****IBS5*
I	Interview	*L55****I****BS5*
B	Before administration of the test	*L55I****B****S5*
A	After administration of the test	*L59I****A****P*
S5	Student	*L55IB****S5***
P	Professor	*L59IA****P***
OF	Observation Format	*L39****OF***
TUAC	Test Usefulness Analysis Chart	*L17****TUAC***

In order to ensure the validity and reliability of the data analysis process, both researchers were present at every single stage of the data collection process, the subsequent transcription, coding, categorization, and analysis of the information. Researchers came to agreements by means of comparing, contrasting, and discussing their insights about the different issues that arose, providing better support to their interpretations of the data collected in light of the study's construct.

Taking into consideration the descriptive nature of this study, the last step was to analyze the data that had been systematized in order to arrive to a narrative description of findings (Fraenkel and Wallen, 2009). These results revolved around the most prominent test usefulness qualities that were evidenced in the analysis of the testing instrument in question, as well as other elements closely related to test design and implementation. With this in mind, the implications of

this research involve aspects that could lead to further reflection and investigation on the language testing practices implemented in this particular context.

4.9. Ethical Considerations

Bearing in mind that this research project entails the participation of human subjects, there were some ethical considerations that had to be addressed. Recalling Mackey and Gass (2005), even though research in the second language field does not represent major risks, approval from the individuals and institutions involved is required. For this particular case, participants include a professor and six students from a Teaching of English undergraduate program. They were invited to be part of this study by means of an informed consent form which explained the purpose and nature of the project, the procedures to be developed, and the generalities and specificities of their participation (See Appendix A). Ethical considerations such as confidentiality, anonymity, voluntary participation, foreseeable risks and effects were explicitly stated in this consent form. Likewise, the course's professor allowed researchers to use the guidelines and rubrics of the test analyzed. All of the information collected will remain confidential and be used exclusively for research purposes; none of the participants' identities will be disclosed; participation in this project was voluntary and did not represent any reasonably foreseeable risk nor negative effects. Apart from signing the consent form, participants were reminded about these considerations when being interviewed and were asked to confirm their willingness to be part of the study. Finally, it is worth mentioning that all of the individuals asked to be part of this project accepted the invitation without any reservation.

5. FINDINGS AND DISCUSSION

5.1 Construct Validity and Authenticity as prominent test usefulness qualities

During the data analysis conducted in this research, it was found that two of the six test usefulness qualities proposed by Bachman and Palmer (1996) were more easily and clearly evidenced than the rest. Elements characteristic of Construct Validity and Authenticity were prominent and allowed to reveal the close relation between both qualities. The harmony between the aspects that comprise these qualities allows to generalize test takers' abilities from test results into the target language use domain. All these elements are described as follows:

5.1.1 Coherence between construct, test task, target language use task, and test taker's language ability as an indicator of construct validity

According to Bachman and Palmer (1996), construct validity is evidenced when there is a coherent relation between the definitions of a language construct, or what is being measured, the test tasks, and the correspondence of said tasks to those developed in a real life scenario, also known as the Target Language Use (TLU) domain. In the case of this study, the test analyzed showed clear signs of a coherent relation between all three of these aspects. This was evidenced as a result of the analysis of the data gathered from two different instruments; those being the interviews conducted with the course's professors and students, and the chart used by the researchers to analyze all data collected.

Table 1 shows excerpts from the interviews previously mentioned that make reference to the three main aspects that comprise construct validity:

Table 1

Correspondence between Construct Validity Aspects: Construct Definition, Test Task, and TLU Task

	Test Usefulness Analysis Chart	Professor Interviews	Test Takers Interviews
Construct Definition	*L10TUAC: The construct definition for this achievement test is based on the course's syllabus and includes the following components:* *- Features of presentations: Body language, eye contact, poise, voice, introduction and closure, and content.* *- Language: Qualitative aspects of spoken language (range, accuracy, fluency, coherence, and interaction), discourse markers, and statistical word choice.*	*L183IBP: It's basically the language from two perspectives: paralinguistic behavior, how the move, stand, speak, see, and the gestures they use in front of an audience, but also the linguistic behavior on the use of discourse markers that are related to the presentation of statistical analysis.*	
Test Task	*L18TUAC: Simulation of a Viva Voce exercise as a rite of academic life including features of presentations and discourse markers.*	*L136IBP: So, they come up with a hypothesis that is driven from a cultural stereotype. Say, for example, "Women are bad drivers", "Religion doesn't fit with humanities but hard sciences", "Students from Mechanical Engineering and Industrial Engineering are party animals". And, so, they had to apply a real test in which they went out, they collected information, they made a survey, they asked real people about it, and, so, they proved their*	*L69IBS1: Es un proyecto acerca de estereotipos culturales, entonces tenemos una parte que es un trabajo escrito y otra que es la sustentación de ese proyecto que hicimos. Entonces, tenemos unas encuestas que hicimos aquí en la universidad, y luego de esas encuestas tuvimos que hacer unas gráficas para tabular como los resultados, y las conclusiones. Entonces ya en grupo realizamos toda esa parte escrita… tuvimos que repartir ciertas cosas para poder hacer la*

		hypotheses if it works or not. And what they do is a presentation using language that is really formal.	*defensa que ya hoy vamos a presentar.*
TLU Task	*L17TUAC: Viva Voce; an oral examination or thesis defense.*	*L173IAP: When I thought of doing this type of task, I referred to an event, which is very common, the Viva Voce event is very common in the academy. It's you having a set or a board of evaluators, you having a thematic, then defending your proposal in front of that.*	*L81IBS1: También vamos a tener que sustentar una tesis, entonces con estos proyectos que son chiquitos pero a la vez tienen una influencia muy grande en uno; no sé, sacar provecho todo lo que pueda de esta presentación. También del feedback que nos den, y ya y ponerlo en práctica el otro semestre en el otro proyecto que tengamos que hacer en las futuras presentaciones orales.*

Regarding the language construct, the definition provided by the professor coincides with the elements described in the analysis chart that was adapted by the researchers. In summary, the construct measured in this test deals with the use of formal register in presentations, which includes linguistic and paralinguistic elements. The linguistic elements measured were the qualitative aspects of spoken language use; being range, accuracy, fluency, coherence, and interaction. The paralinguistic elements, on the other hand, included general features of presentation; being body language, eye contact, voice, poise, and volume. All this gives a clear idea of what was expected from test takers for the implementation of the test.

In the case of the task proposed for the test, there were definitions provided by the course's professor, the researchers, and one of the test takers. Once again, the definition provided in all instances followed the same line. Regarding the test task itself, it can be said that it was the

simulation of a Viva Voce event. This entails the oral defense of a thesis, or a project, to a board of raters that is often developed in academic contexts. In this exam, test takers presented an oral defense of a survey project that was based on cultural stereotypes. To develop the test task, students chose a stereotype, designed a survey, applied it to a target population, tabulated and analyzed the data gathered, and presented the results to an audience of students and a board of evaluators which included the course's professor and 3 guest raters.

The last element addressed in table 1 has to do with the correspondence between the test task and the target language use domain, which is known as the TLU task. Just like with the previous aspects, there was consistency between the definitions of the target language use task provided by the researchers, professor, and test taker. All parties agreed when comparing the test task with a thesis defense, which is a common event in real life academic contexts. This means that the target language use domain was indeed reflected in the task proposed for the test since it was designed to simulate an exercise test takers would have to develop in their lives as students, beyond any particular testing situation.

The coherence between the definitions of all the elements previously mentioned shows clear signs of construct validity within the design of the test analyzed. The construct was specifically defined to measure test takers' abilities regarding the use of formal language in an academic context, and the test task was built upon the same construct definition. Moreover, the task proposed not only measured learners' abilities, but also simulated the target language use task, a Viva Voce event, thus allowing score interpretations to generalize beyond the test and into the target language use domain. All of this is further supported by another key aspect in the consideration of construct validity, being the reflection of test takers' language ability in test

results. The course's professor and test takers were questioned on this matter during the interviews conducted after the test had taken place. To that, they answered the following:

> *L59IAP: It's a sample of what they can do for this specific type of presentations. If you ask them to present an essay or change the format of what they're going to do, I don't think there will be the same results. But since this is a Viva Voce event, which is focalized on how students produce when they are in an academic context, in the line of statistics and ciphers, then, that's what they can do.*
> *L10IAS2: Yes... because, especially in this test, I tried to do it as natural as I can.*
>
> *L40IAS3: Yeah, I think they are real because I should have given a more academic speech instead of a political speech… yeah, I recognize I made some mistakes.*

As seen in the comments made by the course's professor and students, the test's results were considered as real indicators of test takers' languages abilities within the test's context. This became evident when even those students that did not perform at their best, such as student number 3, agreed that the scores they obtained were a reflection of what they could do regarding the construct measured in the test; that is, using formal language in academic presentations. Students showed a positive attitude towards test results since they perceived that the justification for the interpretation of their scores was appropriate, which is further evidence that supports the existence of construct validity in the test. Though this may not comprise test takers' language ability in its entirety, it certainly showed what they are able to do in the test's specific context, which is precisely what was intended to be assessed from the beginning.

As Bachman and Palmer (1996) state, defining construct validity is a very complex issue that is of great value in language testing. The interpretations made upon test scores need to be justified with evidence of the coherent relation between a clearly defined language construct and the tasks proposed in order to develop a given test. Aside from this, and as Frodden, Restrepo, and Maturana (2004) found, one of the main issues regarding construct validity in tests is that most instruments do not provide a clear description of what students are expected to do in order to demonstrate their language ability. That was certainly not the case with the test in question given

that the construct definition was in correspondence with the test task. This suggests that test scores must have been easier to interpret and, ultimately, justify, since there was a clear definition of what was being assessed from the beginning. Furthermore it is necessary to justify that test scores actually reflect test takers' language ability, since this is precisely one of the main parameters that determine the usefulness of tests. Regarding this, it was found that the construct definition was reflected in the test task, and that the latter was designed to simulate the task that should be carried out in the target language use domain. This means that test takers' performance was based on the very same aspects that were being measured, and that it was possible to generalize students' abilities into the target language use domain based on the test's results. All of this provides an evident notion of construct validity in the design of the test itself.

5.1.2 The correspondence between the test task and the TLU task as an indicator of authenticity

In accordance with the test usefulness model proposed by Bachman and Palmer (1996), a higher similarity between test task and TLU task results in a more authentic test. In the test analyzed in this study, there was a high resemblance between both language tasks and, therefore, authenticity was one of its most prominent qualities. This similarity is best evidenced in the following sample from an interview with the course's professor, who is also the test's designer:

> *L69IAP: When I thought of doing this type of task, I referred to an event which is very common. The Viva Voce event is very common in the academy. It's you having a set or a board of evaluators, you having a thematic, then defending your proposal in front of that.*

As mentioned by the professor, the test task is based on a real academic event. As he explained, a Viva Voce is an oral examination in which students are required to defend a thesis in front of an evaluation board. In this sense, the test task consisted on the defense of a project conducted by the test takers. This process was described in detail by one the students interviewed:

> *L55IBS5: We had to create a survey. First, we had to choose a topic. My topic is "women at the wheel". Then, we create a survey and we apply it. For example, in my case, to taxi drivers. Then, we analyze the answers of the questions, and we put them in charts and graphics, such as pie charts, for example. And finally, we explain this project to the audience that we will have.*

In her description, Student 5 explains all of the steps followed in the development of the test task. It involved the selection of a topic, the creation of a survey and its application, the analysis of the answers, and the presentation of the process and results in front of an audience, including an evaluation board. This is quite similar to what a research project actually entails; the formulation of a hypothesis, a data collection and analysis process, a report of findings, and a thesis defense, which would be the so called Viva Voce. This equivalence between test task and TLU task was also revealed in the analysis of the data gathered through the test usefulness analysis chart adapted by the researchers and the observation of the administration of the test:

> *L139TUAC: Do the characteristics of the test tasks correspond to those of the TLU tasks?*
> *Yes, they do. Since the test task consisted of the simulation of a viva voce exercise as a rite of academic life. It included elements of the academic discourse genre known as presentation of projects and the viva voce.*

In the analysis chart, some questions were made regarding each test usefulness quality. In terms of authenticity, the question addressed the correspondence between the characteristics of the test task and those of the TLU domain. As evidenced in *L139TUAC* , the answer to this was positive since the test task consisted of the simulation of a Viva Voce exercise as a rite of academic life. The word "simulation" supports by itself the resemblance to a real-life scenario, and it is included in the test's guidelines as part of the test's objective. Furthermore, this was also observed during the administration of the test, as seen in the next sample taken from the observation format:

> *L39OF: The test task involves the presentation of a survey project through the simulation of a Viva Voce exercise. Students carried out surveys to gather data that was tabulated and is being presented in front of an audience and an evaluation board. The audience is composed of students in the same course that will also present, and the evaluation board counts on the course's professor*

> *and two guest evaluators for the first section and one for the second. There is also some time for questions from the evaluation board at the end of each presentation.*

As it was registered by the researchers, the actual implementation of the test confirmed the similarity between the test task and what students would do in a real Viva Voce event. Apart from this correspondence, another indicator of authenticity is the further relevance of the test task in test takers' lives. Such future application was addressed by the course's professor and the test takers themselves as follows:

> *L69IAP: So, they will use it because they will have to perform a thesis proposal, in which they have to do exactly the same in front of an evaluation board, and they have some statistics, numbers, hypotheses, they conduct some surveys, or they apply, they implement a data collection process, and then when they are there, they have to defend that idea. So, it's precisely the same as they have to do.*

In this sample from one of the interviews with the test designer, he mentioned how students will in fact use their experience with the test when they have to present their own thesis proposals in coming semesters. Additionally, test takers acknowledged the potential relevance of the test in their academic lives with comments such as the one below:

> *L81IBS1: También vamos a tener que sustentar una tesis, entonces con estos proyectos que son chiquitos pero a la vez tienen una influencia muy grande en uno; no sé, sacar provecho todo lo que pueda de esta presentación. También del feedback que nos den, y ya y ponerlo en práctica el otro semestre en el otro proyecto que tengamos que hacer en las futuras presentaciones orales.*

In the previous excerpt, Student 1 highlights the fact that they will have to defend a research project, and that they are likely to take advantage of these types of exercises by implementing the feedback received to improve their performance in future oral presentations.

According to Bachman and Palmer (1996), they "would describe a test task whose characteristics correspond to those of TLU tasks as relatively authentic". This notion is related to the level of equivalence between what students are asked to do in a test and what they are required to do in real life. In the analysis we conducted, this similarity was one of the most evident

characteristics of the test due to the high levels of resemblance in the tasks. After all, a simulation is an imitation of a situation or process, and in this case, the test's objective was to simulate a Viva Voce exercise as a rite of academic life including features of presentations and discourse markers. With this in mind, the importance of authenticity as a quality relies on the possibility to connect test takers' performance with their ability to use the language in the TLU domain; which is actually related to the concept of construct validity we addressed in a previous section. Another key aspect when considering authenticity is how relevant the test task is for students. If students perceive the test as an authentic sample of what they would have to do in their regular contexts, and therefore acknowledging its relevance, they would be more likely to perform at their best.

The fact that authenticity was one of the most prominent qualities in this test analysis, differs from what other researchers in the Colombian context have found regarding this same issue. For instance, Frodden et al. (2004) concluded that "the qualities of assessment that teachers cherish the most are practicality and reliability, and the ones least taken into consideration are authenticity and interactiveness." (p. 190). They attribute the lack of authentic test tasks to teachers' time constraints that lead them to favor reliability and practicality. On the other hand, they argue that designing, developing, and assessing authentic tasks require a great amount of time and effort, which proves to be more beneficial in terms of teaching and learning. In this sense, when students are exposed to tests that resemble real-life situations, their performance becomes authentic as well. Besides, as Bachman and Palmer (1996) state, the way in which learners approach a test depends on how they perceive it. Therefore, if a task is relevant and useful, it will have a positive impact on their ongoing learning process; otherwise, it will be regarded just as another academic requirement.

5.2 The use and implementation of rubrics and grading criteria to assess test takers' performance

One of the key aspects of a test is known as grading criteria, a set of parameters used to measure test takers performance in a more reliable way. It usually includes a description of the aspects that will be evaluated. This not only facilitates the grading process for raters but gives test takers a better understanding of what is expected from them and, later on, of their own results. As Bachman and Palmer (1996) state, "In order for test takers to understand what they are expected to do, and hence to perform at their best, they need to know how their responses are going to be evaluated." (p.189). In this case, the test analyzed counted on two different sets of criteria, one used by the course's professor and the other by the guest evaluators, which were compared in order to obtain test takers' final grades. These rubrics included elements such as features of presentations (body language, eye contact, poise, voice, introduction and closure, and content) and the qualitative aspects of spoken language (range, accuracy, fluency, coherence, and interaction). The complete version of these criteria can be found in the following link: https://goo.gl/N2o6sK .

With this in mind, it was found that there were some clear inconsistencies between the test designer's and test takers' statements about the grading criteria implemented in this test. During the interviews conducted before the administration of the test, the course's professor described the elements to be evaluated. To begin with, this is what he said about the presentation of such criteria:

> *1) L220IBP: "It has already been shown to the students in a previous feedback session I gave them last week, and it has two criteria. Let's call them rubrics."*

As seen in this first sample, when asked if students were provided with the test's grading criteria, the professor assured that it had in fact been presented to them in one of the feedback sessions carried out during the preparation for the test. He also highlighted that two rubrics were

going to be used in the assessment of students' performance; one by him and the other by the professors invited. He described the criteria given to the guest evaluators as follows:

> *2) L224IBP: "It focalizes on elements like the content, the poise, the eye contact. It brings a description of what is expected from the student, and on the other side, the professors make comments about it, and they give a score; a quantitative score to how they performed based on the criteria I gave them."*

The professor's description in the sample above shows the aspects included in the rubric used by the guest evaluators. This grading criteria focused on features of presentations and contained qualitative and quantitative descriptors of the test's levels of achievement, ranging from 2 to 5; 2 being the lowest score possible and 5 being the highest. Just as with this criteria, the course's professor provided an explanation of his own rubric:

> *3) L228IBP: "It includes different elements like the delivery of the student, the language that they use, and the content. Each of them have different subsections which will be ranged from a number 1 to 3; 1, being "poorly executed" or "needs improvement", the other one, "it's sometimes shown", and the other one, "it's really well executed". So, in terms of delivery, we have things like "the speaker used vocal variety to emphasize on different sections of the presentation". So, that's when we use 1, 2, or 3."*

Similar to the guest evaluators' criteria, the rubric used by the professor included descriptions of the aspects to be assessed. In this case, such elements focused on the way students delivered their presentations in terms of language. As shown in L228IBP, students' performances were given a score of 1, 2, or 3, which correspond to "poorly executed", "needs improvement", or "it's sometimes shown" respectively.

In all three samples from the interview with the course's professor, it can be evidenced that there were two rubrics that contained elements related to features of presentations and the qualitative aspects of spoken language. In fact, the guest evaluators' criteria focused on elements such as content, poise, and eye contact, and the professor's focused on the learners' range, accuracy, fluency, coherence, and interaction. Both rubrics used descriptors and levels of

achievement for each aspect being assessed and were later compared to obtain test takers' final grades. Based on samples 1, 2, and 3, it can be said that there was an established set of criteria in order to measure students' performance, and that it was introduced to test takers in advance. Interestingly, when learners were asked about this same matter, they gave a contrasting answer as seen in the comments below:

> *L83IBS2: "They are going to evaluate things such as our body movement, fillers, but I'm not pretty sure because we are used to have the… the rubric? The criteria, but this time I didn't."*

In this sample, Student 2 mentions two of the elements he thought were going to be evaluated; being, body movement, which is one of the features of presentations included in the grading criteria, and the use of fillers, which belongs to the qualitative aspects of spoken language that were also assessed. However, this test taker expressed not to be completely sure about the criteria since he was not provided with a rubric. Another student addressed the same issue by stating that:

> *L163IBS4: "Así como que en, por ejemplo, en Inglés se utiliza mucho que dan como una hojita y entonces le muestran a usted como los puntos o los criterios que el profesor va a tener en cuenta y que uno tiene que meter pues en el speaking, pero por el momento no nos lo han presentado así."*

Similar to L83IBS2, this sample suggests that students were not given a test rubric. In fact, when talking about this matter, Student 4 gave an example from another course in which test takers were usually provided with a piece of paper that includes the criteria to be taken into account by both professor and test takers, and mentioned that up to the moment of the interview, it had not been presented that way.

It is important to mention that the students' comments presented above were taken from the interviews carried out before the administration of the test; and that some of these interviews were even conducted few hours before students' presentations took place. This means that, by

then, the course's professor must have already presented the grading criteria; especially considering that he was interviewed the day before the implementation. Learners' comments lead to believe that, even though they were aware of some of the aspects to be evaluated, they were not presented any type of criteria as opposed to what their professor stated. For instance, sample L83IBS2 shows that elements such as body movement and fillers were going to be taken into account; and these elements were in fact included in the rubrics. Additionally, both comments from samples L83IBS2 and L163IBS4 highlight the fact that test takers were not provided with a rubric; this referred to as a document with the criteria for the evaluation of students' performance. That being said, there is no certainty about the causes of the inconsistency between the course's professor and the students' statements regarding this issue.

Bachman and Palmer (1996) argue that learners' understanding of the aspects to be evaluated can affect their performance. Test takers might have their own notion of what is expected from them, which can shape their responses. Within the test usefulness model proposed by these authors, it becomes necessary to present grading criteria as explicitly as possible, so it will not have a negative impact on test outcomes. The fact that all of the test takers interviewed in this study were not fully aware of the aspects concerning their assessment may have caused them to perform differently from what was expected. Aside from this, even though they acknowledged the importance of having a rubric when stating that they were not provided with one, they did not show strong concerns about the situation. Moreover, the contradiction between the professor's and the students' statements regarding the presentation of the criteria suggests that, at some point, one of the stages of the assessment process presented inconsistencies.

Pineda (2014), concluded that "If teachers are clear about what they have to assess and make it explicit to students, it will impact their teaching practices and also their students' learning.

In addition to this, assessment becomes fairer and more transparent." (p.192). In accordance with Pineda's stance, having an established set of criteria is not enough in itself to ensure a reliable measurement of test takers' abilities if it is not fully understood by them. A way to prevent this from happening is to train students on the use of rubrics as an instrument that not only justifies their results but also allows them to have a better understanding of what is expected from them from the very beginning of the assessment process. This will enable test takers to take advantage of the grading criteria and perform at their best. Besides, it is also useful to involve students in the formulation of the elements to be included in the criteria so as to make the process as democratic as it can be.

5.3 Lack of awareness on test objectives

One of the first steps when designing a test is to set its objectives in order to define what exactly is going to be tested; in other words, what test takers should know or be able to do (Brown, 2003). The objective of the test addressed in this research is to simulate a Viva Voce exercise as a rite of academic life including features of presentations and discourse markers. This information can be found in this link to the test guidelines: https://goo.gl/9odaJr.

After analyzing the data collected, it was revealed that the course's professor and test takers approached the test from a different perspective in terms of its objective. The professor's focus was on the test objective while students focused on the steps they had to follow to develop the test, losing sight of the main goal. The difference in focus is evidenced when comparing their answers to questions related to the test's expected outcomes. In the following comments from the professor's first interview, it is possible to notice the way he conceived the test:

> *L136IBP: What I wanted to do is to provide a space for the students to transform the way they use English into a more formal register, implementing scientific discourse.*

> *L255IBP: I expect that they mix all the elements we've talked about in the course for these twelve weeks. I expect that they don't just consider the last two weeks, but, instead, everything that we've said from the beginning. I expect to see formal students using academic discourse. It won't just be a presentation. It has to be an event where they defend their hypotheses.*

From the samples above, it can be stated that the professor's main objective was to give students an opportunity to use a more academic discourse through the incorporation of all the elements they had studied in the course into the simulation of an academic event called Viva voce. When asked about what he expected students to do, he emphasized on the test as a whole by referring to its main objective. For instance, in *L136IBP* he refers to his intention behind the design of the test, allowing test takers to explore a more formal type of discourse, different to what they normally used. Additionally, in *L255IBP* the professor mentions how he expected students to take advantage of everything they had learned in the course and channel it into the achievement of the test's main objective. Test takers, on the other hand, focused on the process to accomplish that aim as evidenced in their detailed descriptions of what they were expected to do:

> *L55IBS5: We had to create a survey. First, we had to choose a topic. My topic is "women at the wheel". Then, we create a survey and we apply it, for example, in my case, to taxi drivers. Then, we analyze the answers of the questions and we put them in charts and graphics, such as pie charts, for example. And finally, we explain this project to the audience that we will have, and there is another part that is a writing product about the whole project.*

None of the test takers addressed the objective of the test explicitly. Instead, they described the test guidelines step by step as seen in the sample above. All of the students explained how they had to conduct a survey based on a cultural stereotype, consolidate their results, and then present them in front of an audience. They conceived the test as the completion of a series of activities more than the completion of the whole test task, and even though the activities developed depended on each other, they were not addressed as the means to achieve the test's main objective.

Brown (2003) states that it is important to be cautious when implementing an assessment based on a performance since there is a risk of assuming that by doing something for the development of the task, students are actually accomplishing its main objective. He goes on to suggest that teachers ought to set specific and detailed objectives for the performance. The fact that students did not appear to approach the test from a holistic perspective, implies that they may have overlooked its real purpose, and by purpose we mean objective. Even though they followed the steps presented in the guidelines, and in accordance with Brown (2003), developing the activities proposed in a given order is in itself not enough to justify one's performance. Once they developed an activity, they continued with the next one, but in the end it was a collection of actions lacking an ultimate purpose. As happens with the grading criteria, it is fundamental to ensure learners understanding about all the aspects whose misinterpretations could prevent test takers to perform at their best and therefore affect the test's reliability (Bachman & Palmer, 1996). Even more when the test's objective is the starting point and guiding reference for any of the test takers' actions in the development of the test. Surprisingly, the objective of the test analyzed in this study was clearly stated in the guidelines, which leads to believe that test takers focused so much on following each of the instructions on the test task that they lost sight of the test's main goal.

6. CONCLUSIONS

This research project was conceived as a means to delve into the reality of one of the many assessment practices developed in our own English teaching and learning context; this is, as students of a Colombian teaching of English undergraduate program. It focused on the analysis of a language testing instrument designed to measure learners' oral skills in one of the program's courses. This analysis was based on the test usefulness model proposed by Bachman and Palmer (1996), and allowed to determine how its qualities were evidenced in the test. Concludingly, construct validity and authenticity were more prominent than reliability, interactiveness, practicality, and impact. Additionally, elements related to test design and implementation, such as the importance of test objectives and scoring methods, were found.

Regarding construct validity, it was noticed that there was coherence between the construct measured in the test, the test task, the target language use task, and test takers' abilities. The test focused on a specific construct that had been previously addressed as part of the course's contents, and the task proposed for the development of the test was aligned with the construct being measured. Furthermore, the test task resembled a task students would encounter in real life, simulating the demands of the target language use domain. According to Bachman and Palmer (1996), such coherence is what makes possible to justify the interpretations given to test results, which in turn allows to generalize test takers' abilities into the target TLU domain. In this particular case, test's results were perceived as real indicators of language ability by the course's professor and test takers. With this in mind, it was concluded that the test analyzed was designed to measure learners' abilities in regards to a clearly defined language construct, and that it measured very little else; thus making it valid.

In terms of authenticity, the analysis revealed that the test task and the TLU task were similar; this means that what students were asked to do kept a high resemblance to what they would do in a real-life scenario. Furthermore, students themselves acknowledged the future usability and relevance of what they were required to do into their academic and professional lives. As Bachman and Palmer (1996) state, the correspondence between the tasks developed in a test and in real life affects the way in which test takers perform. In this case, the test task consisted of a simulation of an event that students are likely to face in current and future stages of their academic lives and that will probably have a high impact on them. Bearing this in mind, it can be said that the test's resemblance to the TLU domain and the way students responded to such circumstance made this testing experience an authentic exercise.

Construct validity and authenticity are qualities that depend on each other. In this sense, it is not surprising to see them standing together as the most prominent qualities in this study. However, there were other noticeable elements that were not directly connected to the test usefulness model per se; that is the case of test takers' lack of awareness regarding the test's objective and rubric. Setting clear objectives is one of the first and most important steps in the design of a test (Brown, 2004), and even though this was not an issue in the test analyzed, learners were not fully aware of its main goal. Unlike the course's professor, who conceived the test as a whole event with a concrete objective, test takers perceived it as the completion of a set of separate tasks. A similar situation took place in regard to the rubric used to assess students' performance. While the course's professor stated that the test's rubric was presented to learners, they assured not being given any type of document containing the grading criteria. There is no certainty of the reason for such discrepancy between students and professor. Still, the fact that test takers were not

completely sure about how they were going to be assessed could have prevented them from performing at their best (Bachman and Palmer, 1996).

As per the rest of the test usefulness qualities, it is safe to say that although they were not as prominent as construct validity and authenticity, they were indeed evidenced in the analysis conducted. For instance, there were not major inconsistencies in measurement since test takers were all assessed under the same circumstances, which means that the test was reliable; students were required to implement different strategies depending on their individual characteristics in order to develop the task proposed, making the test interactive; all of the resources needed for the test were available and did not present any inconveniences, thus the test can be considered practical; lastly, students acknowledged the current and future relevance of the test in their learning process, and the professor expressed his intention to make improvements to the test based on the overall experience, which is a manifestation of the test's impact.

Different from what Frodden et al. (2004) and Pineda (2014) concluded in their studies, reliability and practicality did not appear as the most favored test usefulness qualities. Even though they were present, they were not prominent enough as to be considered the guiding principles for the design of the test analyzed. The qualities that emerged as the most noticeable were construct validity and authenticity. Instead of aiming solely at consistency of measurement and resource availability, this testing instrument showed, more than anything, coherence between its construct, test task, and real life application. Another interesting outcome was the lack of awareness on language testing from the test takers. Cohen and Fass (2001), Frodden et al. (2004), and Pineda (2014) all concluded that there was a need for teaching training in language assessment practices; however, this was not the case in this study. Rather, it was evidenced that students knowledge about language testing was limited, especially concerning the practices implemented to measure

their own skills and competences. Ultimately, this research project presents another reality of the language testing practices particular to the EFL Colombian scenario, and sets the path for further inquiry in the context it was developed.

6.1 Pedagogical implications

The development of this research project allowed to identify certain pedagogical actions that could enhance the language testing practices implemented in the academic program where it was conducted. These actions involve the creation of a framework to evaluate testing and the instruction of students in this same matter. As it was revealed, the adoption of the test usefulness model proposed by Bachman and Palmer (1996) proved to be of great value when analyzing the testing instrument selected for this study. By having a set of qualities that guided our analysis, it was easier to achieve a better understanding of the test in question. On the other hand, one of the most relevant findings of this project was students' lack of knowledge in regard to language testing. This was evidenced in the way test takers overlooked the test's main objective as well as in their insights regarding the test rubric.

Bearing in mind that testing is inherent to language teaching and learning, it becomes necessary to establish parameters that ensure the appropriateness of the practices implemented. As far as we are aware of, the program in which this research took place does not have a common framework for the evaluation of the tests used to measure learners' abilities. This can result in testing instruments that do not fulfill their intended purpose, that are not aligned with the course's objectives, that do not consider students' individual characteristics, that do not have clear and explicit specifications nor a thorough scoring method, but most of all, that remain inconsistent among and between courses. The analysis exercise developed in this study was based on six main qualities that determine the usefulness of a language test. Even though we were not evaluating the

test but rather describing it in terms of such qualities, we were able to witness the model's applicability to this type of testing situation. It is worth mentioning that the concept of test usefulness relies on the balance of its qualities. This means that all of them need to be taken into account, and that each of them should be granted a certain value depending on the particular test circumstances. This being said, regardless of the model adopted, the pivotal issue is to actually implement an instrument to evaluate the appropriateness of the testing practices within the program.

Aside from establishing an evaluation model for testing, it is fundamental to ensure that all of the parts involved in the assessment process are fully aware of its implications. Frodden et al. (2004) as well as Pineda (2014) highlighted the importance of training teachers in language testing; however, none of them addressed the need for student instruction in this regard. This project allowed us to realize how learners' lack of knowledge about testing affected their interaction with the test. For instance, when students were developing the test task, they did not have a main objective in mind since they did not focus on it in the first place. Also, the fact that they did not have a clear notion of the grading criteria used to assess them could have caused them to perform differently from what was expected. And what is more, they mentioned not being provided with a rubric while the professor assured they were. This leads to believe that if learners had had previous instruction in assessment, they would have been able to face the test in a more conscious way. Whit this in mind, it would be ideal for students to have access to an assessment course that addressed language testing not only from a test designer's perspective but from a test takers' stance. This would help them become more critical towards their learning process, and will eventually influence their testing practices.

The significance of these implications relies on the context where this research project was conducted. For an undergraduate program that instructs future English teachers, it is essential to expose learners to adequate language testing practices since they are likely to replicate them in their future as professionals. Besides, it is also of paramount importance to foster students' metacognition regarding evaluative practices, so their learning process benefits to a greater extent. Apart from providing information about learners' abilities, language tests should be regarded as an opportunity to examine the teaching and learning practices implemented in a given course or program. Thus, language testing is not a process to be taken lightly but rather carefully.

6.2 Research implications

The development of this research project also revealed the need for further inquiry on the language testing practices implemented in this particular EFL context. In fact, prior to this study, there was no record of any investigations regarding this issue. Taking this into consideration, it is suggested that the program encourages the analysis of the practices used to assess learners' skills and competences by means of formal research.

The present study can be regarded as an initial step towards the consolidation of this investigative endeavor. It is important to highlight that this project focused on a single testing instrument from one of the program's courses, and that it measured only one of four language skills. It would be interesting to have a more complete notion of the current testing practices developed in the program. This would allow to identify the strengths and weaknesses of the testing practices conducted in each course to eventually take action towards their enhancement. It would also raise awareness on the importance of language testing and would consolidate its role in the academic program. Furthermore, it would promote professors' reflection about the appropriateness of their testing practices, as well as students' critical thinking regarding their assessment. Apart

from filling the current program's research gap regarding testing, these actions would be have a positive impact on language teaching and learning in the established context.

7. RESEARCH LIMITATIONS

During the implementation of this research project, there were a series of circumstances that affected the way it was conducted. First of all, the interpretations of the data collected were subject to the researchers' own understanding of the theoretical concepts that support the study. On the other hand, student participants' lack of knowledge regarding language testing hindered the data collection and analysis process. One of the challenges often faced consisted in deciphering what students were trying to convey as well as explaining testing concepts since they were not familiar with the particular metalanguage used in this study. Regardless of these limitations, this research process did not entail any major inconveniences.

8. REFERENCES

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, D. H. (2004). *Language assessment: Principles and classroom practices.* New York, NY: Longman.

Carr, N. (2011). *Designing and Analyzing Language Tests*. Oxford: Oxford University Press.

Cohen, A., & Fass, L. (2001). Oral language instructions: teacher and learner beliefs and the reality in EFL classes at a colombian university. *Íkala, revista de lenguaje y cultura*, 6(11-12), 43-62.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP.

Fraenkel, J. & Wallen, N. (2007) *How to Design and Evaluate Research in Education.* (7th Ed.) San Francisco, CA: McGraw-Hill.

Frodden, M., Restrepo, M., & Maturana, L. (2004). Analysis of assessment instruments used in foreign language teaching. *Íkala, revista de lenguaje y cultura*, 9(15), 170-201.

Fulcher, G. (2014). *Testing Second Language Speaking*. USA: *Routledge*.

Luoma, S. (2004). *Assessing Speaking*. UK: Cambridge University Press.

Merriam, S. (2009). *Qualitative Research: A Guide to Design and Implementation.* San Francisco, CA: Jossey-Bass.

Pineda, D. (2014). The feasibility of assessing teenagers' oral English Language performance with a rubric. *PROFILE Issues in Teachers' Professional Development*, 16(1), 181-198.

Appendix A

Consent Form

Consent to Participate in a Research Study
(Adapted from Mackey and Gass, 2005, p. 33)
Licenciatura en Lengua Inglesa
Universidad Tecnológica de Pereira

Title of Study:
Analysis of an Oral Skills Testing Instrument in a Teaching of English Undergraduate Program

Researchers:
Name: John Edison Carvajal Arenas **Phone:** 3106546265
E-mail: jcarvajalarenas@utp.edu.co

Name: Luisa Fernanda Castaño Ocampo **Phone:** 314 609 0454
E-mail: LFCO1990@utp.edu.co

Introduction:

You are invited to consider participating in this research study. We will analyze one of the instruments used to test oral skills from the Academic Discourse I course. This form will describe the purpose and nature of the study and your rights as a participant in the study. The decision to participate or not is yours. If you decide to participate, please sign and date the last line of this form.

Explanation of the study:

We will be analyzing a test used to measure students' oral abilities in the Academic Discourse I course, based on six qualities that determine its usefulness: Reliability, construct validity, authenticity, interactiveness, practicality, and impact. In order to do so, six students enrolled in the course as well as the course's professor will participate in this research project. As part of the study, researchers will conduct observations that will take place during the implementation of the testing instrument. Additionally, the course's students and professor will meet with the researchers for individual oral interviews before and after the test is implemented. Interviews will be audio-recorded.

Confidentiality:

All of the information collected will be confidential and only used for research purposes. This means that your identity will be anonymous; in other words, no one besides the researchers will know your name. Whenever data from this study are published, your name will not be used. The data will be stored on a computer, and only the researchers will have access to it.

Your participation:

Participating in this study is strictly voluntary and does not imply any reasonably foreseeable risks nor negative effects. If at any point you change your mind and no longer want to participate, you are free to let researchers know about your decision. You will not be paid for participating in this study. If you have any questions about the research, you can contact the study's researchers by phone at 314 609 0454, by e-mail at aosti.rp@gmail.com, or in person at Room 12-402 of the Humanities and Fine Arts Faculty.

Researchers' statement

We have fully explained this study to the participants. We have discussed the activities and have answered all of the questions that the participants asked.

Researchers' signatures

John Edison Carvajal Arenas	**Luisa Fernanda Castaño Ocampo**

Date: _____

Participants' consent:

I have read the information provided in this Informed Consent Form. All my questions were answered to my satisfaction. I voluntarily agree to participate in this study.

Participants' name	**Participants' signature**	**Date**

Appendix B

Test Guidelines

THE SURVEY PROJECT

Partial Test 2

The following document describes the procedure of the second partial test which is based on the development of an academic genre in discourse called oral presentation of projects and the viva voce. Please follow these guidelines to complete the assignment.

Time of presentations: Wednesday 26 of October, 2016.
- **Group 1: From 16.00 to 17.30 p.m.**
- **Group 2: From 17.45 to 19.00 p.m.**

Objective: Simulate a Viva Voce exercise as a rite of academic life including features of presentations and discourse markers.
Instructions: For the development of this exercise you must present 2 products which are described below, use the criteria to identify what you need to submit.

PRODUCT 1: Written projects.

All presentations must be supported with a written document that contains:

- Cover Page with title, names of the members of the group, subject, faculty, program and date.
- Objective of the survey.
- Procedure of data collection.
- Data tabulation and charts.
- Results and conclusions.
- Appendix with a sample of the survey format.

PRODUCT 2: Oral defence of projects. Presentations will be developed in room 13B - 414 and the group will be divided in two. Some outside evaluators will attend to the session. Each groups has a maximum of 15 minutes to present their projects. All members of the group must be part of the presentation.

Consider this anatomy for a successful presentation:

Greetings
(Personal intro)
Presentation of candidates
Outline of the presentation
Objective of the survey
Procedure
Data tabulation
Conclusions
Questions from evaluation board

Important Note: Slides, PPTs, or keynotes must be sent the latest on Tuesday 25th of October. 10:00 p.m. No Pendrives or USBs will be accepted in the presentation. Please keep this in mind for the preparation and execution of your project presentation.

Appendix C1

Professor's Test Rubric

SURVEY PROJECT

EVALUATION RUBRIC

Name of the presenter:

3= strong 2= manifested in some cases 1= needs improvement

Delivery

The speaker seemed comfortable executing natural speech.	1 2 3
The speaker made eye contact with the audience.	1 2 3
The speaker was fluent.	1 2 3
The speaker did not add extra, unnecessary words like "umm" and "you know."	1 2 3
The speaker's posture was appropriate.	1 2 3
If the speaker used gestures, they were effective.	1 2 3
The speaker used appropriate pitch.	1 2 3
The speaker used appropriate volume.	1 2 3
The speaker used vocal variety.	1 2 3

Language

The speaker used discourse markers when appropriate.	1 2 3
The speaker used good statistical word choice.	1 2 3
The speaker's speech was easy to follow and understand.	1 2 3
The speaker used correct grammar.	1 2 3

Timing

The speaker spoke in the allowable 5- to 7-minute time frame.	1 2 3

Comments:

_____

_____

_____

Pronunciation slips:

Appendix C2

Guest Evaluators' Test Rubric

Survey presentation rubric

For evaluator's use

	5	4	3	2
Body Language	Movements are fluid and help the audience visualize the report	Movements and/or gestures are neutral in affecting presentation	Very little movement or descriptive gestures OR Movement or gestures are somewhat distracting	No movement or descriptive gestures OR Movement and/or descriptive gestures are distracting and take away from presentation
Eye Contact	Holds attention of entire audience with the use of direct and appropriate eye contact	Fairly consistent use of direct eye contact with most of the audience	Displays minimal eye contact with all of the audience OR Focuses on only 1-2 people	No eye contact with audience
Introduction and Closure	Student delivers open and closing remarks that capture the attention of the audience and set the mood	Introductory and closing remarks are clearly delivered	Student clearly uses either an introductory or closing remark but not both	Student does not display clear introductory or closing remarks
Poise	Student appears relaxed and self confident, makes no mistakes in articulation or	Makes minor mistakes but quickly recovers from them; displays little tension	Displays mild tension; has trouble recovering from mistakes	Tension and nervousness are obvious; has trouble recovering from mistakes
Voice	Use of fluid speech and inflection maintains the interest of the audience	Satisfactory use of inflection, inconsistent use of fluid speech	Displays some level of voice inflection throughout delivery	Consistently speaks in a monotone voice
Content	Content is well thought out; arguments are clear and well understood	Content is organized; there is some confusion in speech	Content is not well-organized but major topics are addressed	Content is disorganized; does not address the topic at hand

Adapted from:

Group	Name of presenter	Body language	Eye contact	Intro and closure	Poise	Voice	Content	Overall
1	1.							
	2.							
	3.							
	4.							
2	5.							
	6.							
	7.							
	8.							
3	9.							
	10.							
	11.							
4	12.							
	13.							
	14.							
	15.							
5	16.							
	17.							
	18.							
	19.							

In case of having extra comments write the number of the student and the comment next to it.

Appendix D

Test Usefulness Analysis Chart

Test Usefulness Analysis Chart
(Adapted from Bachman And Palmer, 1996)

Description of the Test

Purpose of the Test

Construct to be measured

Description of the task

TLU Task:

Test Task:

Characteristics of the Setting

Physical Setting:

Participants:

Time:

Characteristics of the Test Rubric

Instructions:

Time Allotment:

Scoring Method:

Characteristics of the Input

Characteristics of the Expected Response

Test Usefulness Qualities

Reliability

- Does the setting vary?
- Does the test rubric vary?
- Does the input vary?
- Does the expected response?

Construct Validity

- Is the construct for the test clearly defined?
- Is the construct for the test relevant to the purpose of the test?
- Does the test task reflect the construct definition?
- Do the scoring procedures reflect the construct definition?
- Is the test setting likely to cause different test takers to perform different?
- Possible sources of bias in the task characteristics:
    *Setting
    *Rubric
    *Input
    *Expected Response

Interactiveness

- Does the test task presuppose the appropriate area or level of topical knowledge?
- Can test takers be expected to have this area or level of topical knowledge?
- Are the characteristics of the test task suitable for test takers?
- Does the test task involve a narrow or a wide range of areas of language knowledge?
- What language functions other than the simple demonstration of language ability are involved in the task?
- Are the test tasks independent?
- Is there opportunity for strategy involvement?
- Is the test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best?

Authenticity

- Do the characteristics of the test tasks correspond to those of the TLU tasks?
Practicality
- What type and amounts of resources are required to develop the test? - Are the resources required to develop the test available?
Impact
- Is the feedback provided to test takers relevant, complete, and meaningful? - Are decisions, procedures, and criteria applied uniformly to all groups of test takers? - Are test takers fully informed about the test procedures and criteria? - Are these procedures and criteria actually followed in making decisions? - Are the characteristics of the test consistent with the characteristics of teaching and learning activities? - Is the purpose of the test consistent with the values and goals of the course?

Appendix E1

Professor Interview Before Test Implementation

Professor Interview Protocol I

Research Project: Analysis of an oral skills testing instrument in a teaching of English undergraduate program

Date: _____ Time: _____ Location: _____

Interviewers: _____

Interviewee: _____

- **Introduction**

The current interview is part of the data collection process developed in this research project, whose purpose is to analyze one of the testing instruments implemented in the Academic Discourse I course from the Teaching of English undergraduate program at Universidad Tecnológica de Pereira. It is of high value to consider the course's professor and students insights before and after the implementation of the test to be analyzed.

You have been invited to voluntarily participate in this project. All of the information collected will be confidential and exclusively used for research purposes.

Thank you for your participation.

- **Interview Questions**

A. Interviewee background:

1. How long have you been a professor at the Teaching of English undergraduate program?

2. What courses have you guided in the program?

3. Do you have any academic or professional experience in language testing?

B. Language Testing

1. What is language testing for you?

2. Do you follow any language testing principles?

Probes: If so, which are they?

3. What are the main aspects that you take into consideration when designing a language test?

4. Are you familiar with the concept of Test Usefulness?

Probes: If so, can you explain it briefly?

C. Academic Discourse I Course

1. How long have you been guiding the Academic Discourse I course?

2. Are there any particular aspects that you take into considerations when designing language tests for the course?

Probes: If so, which are they?

3. Can you describe the evaluation and assessment scheme of the course?

Probes: How many tests are implemented? What type of tests are implemented? How are they graded?

D. Academic Discourse I Course's 2nd Partial Test

1. What are the main aspects you took into consideration when designing this test?

Probes: Were students involved in the process at all? How?

2. What is the purpose of this test?

3. What are students expected to do?

Probes: In terms of language ability? Content? Are they provided with any guidelines?

4. What teaching and learning activities were developed to prepare students for the test?

5. What are the criteria to measure students' performance?

Probes: Are all aspects from the grading criteria worth the same? Do students have access to such criteria?

6. Will students be provided with feedback?

7. What are your expectations regarding the test and students' performance?

- **Closure**

1. Thank you to interviewee

2. Reassure confidentiality

3. 2nd interview reminder

Post Interview Comments and/or Observations:

_____
_____
_____
_____
_____
_____
_____

Appendix E2

Student Interview Before Test Implementation

Student Interview Protocol I

Research Project: Analysis of an oral skills testing instrument in a teaching of English undergraduate program

Date: _____ Time: _____ Location: _____

Interviewers: _____

Interviewee: _____

- **Introduction**

The current interview is part of the data collection process developed in this research project, whose purpose is to analyze one of the testing instruments implemented in the Academic Discourse I course from the Teaching of English undergraduate program at Universidad Tecnológica de Pereira. It is of high value to consider the course's professor and students insights before and after the implementation of the test to be analyzed.

You have been invited to voluntarily participate in this project. All of the information collected will be confidential and exclusively used for research purposes.

Thank you for your participation.

- **Interview Questions**

A. Interviewee background:

1. How long have you been a student at the Teaching of English undergraduate program?

2. In which semester are you enrolled?

B. Language Testing

1. Do you think tests are necessary?

2. What do you do before presenting a test?

3. How do you feel before and after taking a test?

4. How do you feel about speaking tests?

5. What do you do before for a speaking test?

C. Academic Discourse I Course

1. How many times have you taken the Academic Discourse I course?

Probes: If you have taken the course more than once, please describe your previous experience(s) briefly.

2. Can you describe your experience in this Academic Discourse I course?

3. Are you familiar with the course's evaluation and assessment scheme? If so, can you describe it?

Probes: How many partial tests does it have? What percentages of the final grade do they account for? What do they consist of?

4. What do you think about the evaluation and assessment scheme proposed for the course?

5. Can you briefly describe your experience with the first partial test of the course?

Probes: How did you do? How did you feel before and after taking the test? What were you expected to do?

D. Academic Discourse I Course's 2nd Partial Test

1. How do you feel about the 2nd partial test of the course?

2. What are you expected to do for the 2nd partial test of the course?

Probes: Were you provided with any guidelines?

3. Have you prepared for the test? How?

Probes: Did you have advisory sessions? How much time were you given? What resources do you need for this test? Are they available? Have you had any difficulties with this?

4. Are you familiar with the test's grading criteria?

Probes: Can you describe it?

5. What are your expectations regarding your performance in the test?

- **Closure**

1. Thank you to interviewee

2. Reassure confidentiality

3. 2nd interview reminder

Post Interview Comments and/or Observations:

_____

_____

_____

***Note:** When necessary, questions to students will be asked in both English and Spanish, their mother tongue, to avoid misunderstandings. Students' answers will be recorded in their first language to facilitate the interview process.

Appendix F

Observation Format

Analysis of an Oral Skills Testing Instrument in a Teaching of English Undergraduate Program **Licenciatura en Lengua Inglesa / Universidad Tecnológica de Pereira** **Observation Format**		
Date:	**Time:**	**Observers:**
P: Professor; **S1:** Student 1; **S2:** Student 2; **S3:** Student 3; **S4:** Student 4; **S5:** Student 5; **S6:** Student 6		
Testing Instrument: The Survey Project - Viva Voce (Project's oral defense)		
Test Usefulness Quality		**Observation Comments**
Reliability (Consistency of measurement): A reliable test score will be consistent across different characteristics of the testing situation. Setting: resources, time, turn, environment, audience, rater, evaluators, etc. (Setting, rubric, input, expected response)		
Construct Validity (Meaningfulness and appropriateness of the interpretations made on the basis of test scores): Language ability construct; purpose of the test (simulate a Viva Voce exercise as a rite of academic life including features of presentations and discourse markers); grading criteria; task; possible sources of bias (setting, rubric, input, and expected response)		
Authenticity (The degree of correspondence of the characteristics of a given language test task to the features of a TLU task): Test task (Simulation) and TLU task (Viva Voce exercise as a rite of academic life)		
Interactiveness (The extent and type of involvement of the test takers' individual characteristics in accomplishing a test task): Area and level of topical knowledge, test takers' behavior, areas of language knowledge, language function, test tasks interdependence, strategy involvement, affective responses.		
Practicality (Availability of resources for the design, development, and implementation of the test): Human resources, material resources, and time.		
Impact (The effects of the test on the individuals involved): Test takers' language use; test takers' involvement; feedback; rubric; grading criteria; purpose of the test; test takers' behavior.		

Appendix G1

Professor Interview After Test Implementation

Professor Interview Protocol II

Research Project: Analysis of an oral skills testing instrument in a teaching of English undergraduate program

Date: _____ Time: _____ Location: _____

Interviewers: _____

Interviewee: _____

- **Introduction**

The current interview is part of the data collection process developed in this research project, whose purpose is to analyze one of the testing instruments implemented in the Academic Discourse I course from the Teaching of English undergraduate program at Universidad Tecnológica de Pereira. It is of high value to consider the course's professor and students insights before and after the implementation of the test to be analyzed.

You have been invited to voluntarily participate in this project. All of the information collected will be confidential and exclusively used for research purposes.

Thank you for your participation.

- **Interview Questions**

A. Academic Discourse I Course's 2nd Partial Test

1. Were you satisfied with the test's results? Why?

Probes: What about students' performance? Did they do what they were expected to do?

B. Test Usefulness Qualities

- **Reliability**

2. What was students' average final grade in the test?

Probe: Were there any major inconsistencies among students' grades? Why do you think this happened? Did the fact of having two or three evaluators affect students' final grades?

- **Validity**

3. Would you say students' results are indicators of their real language ability? Why?

- **Authenticity**

4. Do you consider that the test task reflected what students would have to do in a real-life scenario? Why?

Probes: Will this experience be useful in other situations? any future occasions?

- **Interactiveness**

5. Are you aware of students having any type of difficulties while relating or interacting with the test?

Probes: Regarding the task, the grading criteria, the resources needed, etc.

- **Practicality**

6. Were all the resources required for the development of the test available?

Probes: What about in the case of students?

- **Impact**

7. What was the impact that the test had on students and on the course?

Probes: How did students react to the test's results? Did they receive feedback? How did they react to it? Based on the test's results, are you planning to make any modifications to the test or the course itself? Why?

- **Closure**

1. Thank you to interviewee

2. Reassure confidentiality

Post Interview Comments and/or Observations:

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Appendix G2

Student Interview After Test Implementation

Student Interview Protocol II

Research Project: Analysis of an oral skills testing instrument in a teaching of English undergraduate program

Date: _____ Time: _____ Location: _____

Interviewers: _____

Interviewee: _____

- **Introduction**

The current interview is part of the data collection process developed in this research project, whose purpose is to analyze one of the testing instruments implemented in the Academic Discourse I course from the Teaching of English undergraduate program at Universidad Tecnológica de Pereira. It is of high value to consider the course's professor and students insights before and after the implementation of the test to be analyzed.

You have been invited to voluntarily participate in this project. All of the information collected will be confidential and exclusively used for research purposes.

Thank you for your participation.

- **Interview Questions**

A. Academic Discourse I Course's 2nd Partial Test

1. Were you satisfied with the test's results? Why?

Probes: How was your performance?

B. Test Usefulness Qualities

- **Reliability**

2. What was your final grade for the test? What was the final grade for the written and oral products?

Probe: Were you presented with the grades that all evaluators gave?

- **Validity**

3. Would you say that your result is an indicator of your real language ability? Why?

- **Authenticity**

4. Do you consider that what you had to do in the test is similar to what you would have to do in a real-life scenario? Why?

Probes: Do you think this experience is going to be useful in the future? Why?

- **Interactiveness**

5. Did you have any difficulties interacting with the test?

Probes: Regarding the task, the grading criteria, the resources needed, etc.

- **Practicality**

6. Were all the resources required for the development of the test available?

- **Impact**

7. What was the impact that the test had on you?

Probes: How did you react to the test's results? Did you receive feedback? From whom? How did you react to the feedback? Based on the test's results, are you planning to follow the suggestions you were given? Which ones?

- **Closure**

1. Thank you to interviewee

2. Reassure confidentiality

Post Interview Comments and/or Observations:

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____